基于关联数据的类簇语义揭示模型研究

崔家旺 [1, 2]　李春旺 [1]
[1](中国科学院文献情报中心　北京 100190)
[2](中国科学院大学　北京 100049)

摘要:【目的】调研基于关联数据揭示类簇内主题词间语义关系的模型和技术方法。【方法】利用 Google Scholar、Springer、CNKI 等检索与研究主题相关的文献, 调研分析并梳理当前类簇分析和语义关系揭示相关研究, 构建基于关联数据的类簇语义关系揭示模型, 通过实验验证模型的有效性。【结果】实验结果表明, 利用关联数据可以有效揭示主题词间语义关系, 弥补传统共词聚类分析在语义方面的不足。【局限】受实验数据限制, 目前揭示出的语义关系局限于上下位类关系、类与实例关系和相关关系等类型, 未考虑关联数据质量问题对语义揭示结果造成的影响。【结论】提出的基于关联数据的类簇语义关系揭示模型可以有效揭示主题词间语义关系, 为共词聚类结果的理解和分析提供一种新的方式。

关键词: 关联数据　共词聚类　类簇　语义揭示模型
分类号: G25

1　引　言

共词聚类分析根据物以类聚的原理将本身没有类别的主题词聚集成代表不同研究子领域的类簇, 通过分析这些类簇可以清晰直观地揭示学科的主题结构与变化[1]。根据聚类原理, 类簇将距离最短的主题词聚集在一起而未考虑词间的逻辑关系, 这样造成的后果是类簇因缺少主题词间的语义关系而难以理解。关联数据的发布与应用为共词聚类研究的发展提供了新契机, 特别是关联数据预先建立了大量权威、准确的属性关系, 每个数据对象包括多种属性和特征, 从而为实现跨学科领域、跨数据源的精准语义关系揭示提供有效支撑。

2　相关研究概述

类簇分析从分析层次上可分为紧密度分析和语义关系揭示两种。类簇的紧密度分析主要衡量聚类的紧密程度, 相关研究主要包括粘合力、密度等类簇分析指标以及共词聚类与其他辅助方法的结合。类簇的语义关系揭示主要从知识发现的角度探索类簇内部语义关系, 相关研究主要包括: 学科专家参与、共词关联分析、文本挖掘、基于本体和词表、基于关联数据的方法等。

(1) 学科专家参与, 张树良等[2]提出共词聚类的过程应有学科专家的介入, 学科专家通过人工梳理的方式帮助人们理解类簇内和类簇间的语义关系, 弥补了共词聚类对数学统计的依赖。

(2) 共词关联分析, 关联规则是描述一个事物中物品之间同时出现的规律的知识模式, 共词关联分析以此为原理, 通过关联统计方法揭示主题词间的依存关系。张晗等[3]利用关联规则算法对 4 种抗肿瘤药物主题词和副主题词组配模式进行分析, 抽取出与这 4 类药物有关的、有效的语义关系搭配模式。张晗等[4]根据书目文献数据库中主题词/副主题词之间的语义关联规则抽取知识, 获得具体的药物与疾病之间的知识。Cimino 等[5]对主题词和副主题词的组配规则进行研究, 通过使用简单的模式匹配规则来自动生成医学概念之间的语义关系。

(3) 文本挖掘, 面向语义关系发现的文本挖掘主要通过对 NLP 进行扫描和自动化处理, 发现概念术语

通讯作者:崔家旺, ORCID: 0000-0002-1596-2195, E-mail: cuijiawang@mail.las.ac.cn。

及概念术语间存在的语义关系。刘明岩[6]结合文本挖掘和本体自动构建的方法探索了军用飞机领域概念间的语义关系。

(4) 基于本体和词表的语义关系发现主要从已知的概念间的语义关系出发。张小刚[7]结合概率论在中医药语言系统的应用基础上，利用关联关系分布推断中医药领域未知的语义关系类型。魏来[8]以词表为语义基础，引入关联词典机制，通过识别标签集中的标签同在线词表概念体系之间的关系，进而识别出标签之间的语义关系。

(5) 基于关联数据的语义关系发现研究还处于探索阶段。Tiddi 等[9]提出的 Dedalo 启发式关联数据遍历挖掘系统具有一定代表性，Dedalo 通过启发式的迭代检索关联数据寻找簇内实体间共同路径，进而形成簇内实体共有的语义关系。Taheriyan 等[10]通过语义标注和构建语义关联的方式利用关联数据推断结构化资源的语义关系。此外，还有一些关联数据挖掘相关技术对本研究有重要借鉴意义。在国内，李楠等[11]、李俊等[12]分别总结了基于关联数据的数据挖掘相关研究，提出基于关联数据的知识发现模型。高劲松等[13]在关联数据的知识发现过程金字塔的基础上提出基于关联数据的知识发现模型。宋丽娜[14]提出关联数据环境下基于知识地图的隐性知识发现模型。刘龙[15]提出基于关联数据的知识发现过程模型。与国内相比，国外研究较为丰富。Narasimha 等[16]提出的 LiDDM 关联数据挖掘系统及 Paulheim 等[17]提出的 FeGeLOD 特征提取器通过格式转化或特征提取将关联数据转化为适合传统数据挖掘算法的格式。Ramezani 等[18]提出的 SWApriori 和 Personeni 等[19]提出的 ILP 学习方法通过改进传统数据挖掘算法将其应用于 RDF 格式数据进行关联数据的挖掘。Jiang 等[20]提出的频繁子图挖掘方法及 Li 等[21]提出的深度学习方法针对关联数据的属性链和节点等结构信息进行挖掘。

每种方法都有一定的缺陷，专家参与方法的缺陷在于成本高、难以推广；基于关联分析的语义关系发现的缺陷在于只能发现某些特定类型的语义关系；基于文本挖掘的方法缺陷在于文本语料库通常缺乏足够的结构化信息，本体和词表的结构严谨但覆盖程度和语义关联程度交叉不足，许多本体和词表相关往往在大小和规模上有所限制，难以覆盖到足够丰富的概念以及概念之间的关系。关联数据作为一个可供语义挖掘的重要资源，在规模和结构上体现出双重优势，因此基于关联数据揭示类簇语义关系虽然属于一种新的尝试，但伴随着 LOD 数据资源和相关技术的快速发展，这种新的语义关系揭示方法可能会成为未来研究发展的一个趋势。

3 基于关联数据的类簇语义揭示模型

类簇内的主题词对应关联数据中的节点，根据关联数据的网状结构特征，主题词节点间最大距离为 3 可能存在的关联关系如图 1 所示。

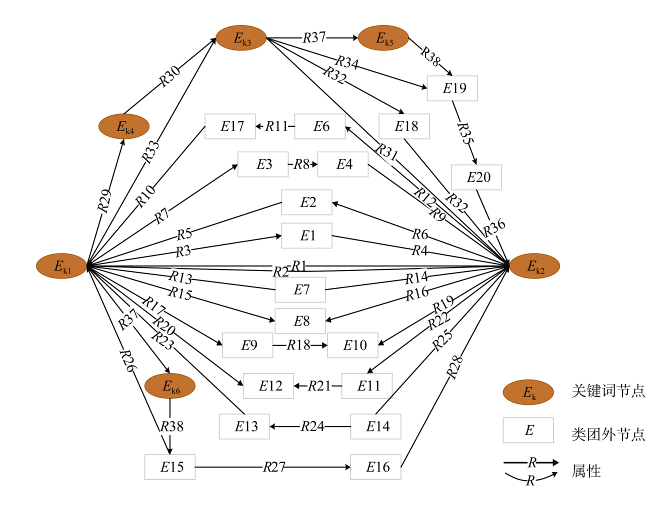


图 1 主题词节点间关联关系示意图

其中棕色椭圆框(E_k)表示类簇内主题词对应关联数据中的节点，即主题词节点；白色方框代表的类簇外节点(E)是指通过关联数据挖掘发现的新节点，节点间的直线/曲线代表属性关系(R)。本文将研究范围限定在主题词距离不超过 3 的关联关系，原因如下:

(1) 最大距离为 3 保证了充足的关联关系。LOD 是典型的小世界图，这种图的特点就是无论网络规模多大，一般搜索路径的最大步数是一个比较稳定的值，研究表明 LOD 中节点间平均最短路径长度为 2.4[22];

(2) 根据路径综合重要性评价方法，距离较远的关联关系重要性较低，缺乏语义揭示的价值;

(3) 关联数据图挖掘的检索空间呈指数上升，更长的路径会导致更大的时间开销。

3.1 基于关联数据的类簇语义揭示模型结构

为准确描述类簇内关联关系，本文提出以下定义:

(1) 关联数据图: 关联数据图是由 RDF 数据构成

的有向图, 图中节点是由 URI 标注的主语或对象, 边是一组具有 URI 标注的属性。

(2) 关联路径: 本文将从主题词节点 E_{k1} 出发到主题词节点 E_{k2} 之间所经过的属性 R 和节点 E 的集合定义为关联路径, 从 E_{k1} 经过节点 $E1$ 到 E_{k2} 的一条关联路径可以表示为: $E_{k1} \xrightarrow{R1} E1 \xrightarrow{R2} E_{k2}$, 其中 E_{k1} 、E_{k2} 表示主题词节点, E1 为关联数据挖掘发现的类簇外节点, $R1$ 和 $R2$ 表示节点间属性关系。关联路径的长度指路径拥有的属性数量, 例如: $E_{k1} \xrightarrow{R1} E1 \xrightarrow{R2} E_{k2}$ 就是一条长度等于 2 的关联路径;

(3) 路径和属性方向: 从主题词 E_{k1} 到主题词 E_{k2} 的关联路径的方向表示为 $E_{k1} \longrightarrow E_{k2}$, 关联路径中的属性关系方向与 $E_{k1} \longrightarrow E_{k2}$ 相同的为正向属性, 属性关系与关联路径方向相反则为逆向属性。例如, 在关联路径 $E_{k1} \xleftarrow{R1} E1 \xrightarrow{R2} E_{k2}$ 中, $\xleftarrow{R1}$ 为逆向属性, $\xrightarrow{R2}$ 为正向属性。

3.2　关联路径分类

由于多个主题词节点之间的关联路径错综复杂难以理解, 本文从两两主题词节点间的语义关系出发, 逐步探索整个类簇内主题词之间的语义关系。以图 1 中主题词节点 E_{k1} 和 E_{k2} 为例, 根据关联路径长度和属性方向的不同, E_{k1} 和 E_{k2} 间的关联路径可分为: 直接关联、间接关联、最近公共祖先节点关联、最近公共子孙节点关联等 4 类, 不同类型的关联路径对应不同类型语义关系。

(1) 直接关联(Direct Relation, DR); 直接关联指的是主题词节点间长度为 1 的关联路径, 主题词节点 E_{k1} 和 E_{k2} 间存在 $E_{k1} \xrightarrow{R1} E_{k2}$ 、$E_{k1} \xleftarrow{R2} E_{k2}$ 两种直接关联。

(2) 间接关联(Indirect Relation, IR); 间接关联指主题词间长度大于等于 2 且不存在逆向属性的关联路径。如图 2 所示, 主题词 E_{k1} 和 E_{k2} 之间长度为 2 的间接关联有 $E_{k1} \xrightarrow{R3} E1 \xrightarrow{R4} E_{k2}$ 和 $E_{k2} \xrightarrow{R6} E2 \xrightarrow{R7} E_{k1}$ 等两种, 关联路径长度为 3 时存在 $E_{k1} \xrightarrow{R7} E3 \xrightarrow{R8} E4 \xrightarrow{R9} E_{k2}$ 和 $E_{k2} \xrightarrow{R12} E6 \xrightarrow{R11} E17 \xrightarrow{R10} E_{k1}$ 等两种间接关联。

(3) 最近公共祖先节点关联; 最近公共祖先

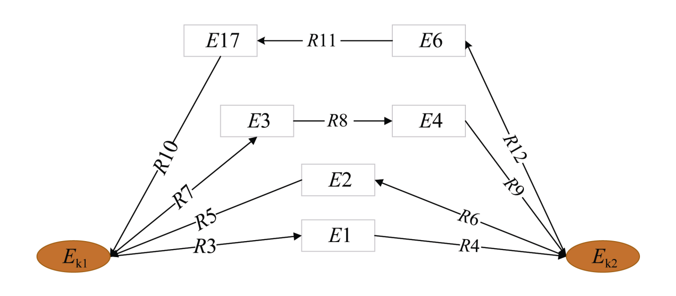


图 2　间接关联示意图

(Lowest Common Ancestor, LCA)的定义为: 对于有根树 T 的两个节点 u、v, 最近公共祖先 LCA(T, u, v)表示一个节点 x, 满足 x 是节点 u 和节点 v 的祖先且 x 的深度尽可能大。在关联数据中也存在类似的结构, 存在最近公共祖先节点的关联路径被定义为最近公共祖先节点关联(Lowest Common Ancestor Relation, LCAR)。最近公共关联祖先关联的定义如下: 通过最短的属性链向两个主题词节点的节点被称作主题词的最近公共祖先节点。如图 3 所示, 当关联路径长度为 2 时, 存在 $E_{k1} \xleftarrow{R13} E7(LCA) \xrightarrow{R14} E_{k2}$ 一种 LCAR。当关联路径长度为 3 时, 存在 $E_{k1} \xleftarrow{R23} E13 \xleftarrow{R24} E14(LCA) \xrightarrow{R25} E_{k2}$ 和 $E_{k1} \xleftarrow{R26} E15(LCA) \xrightarrow{R27} E16 \xrightarrow{R28} E_{k2}$ 两种 LCAR。

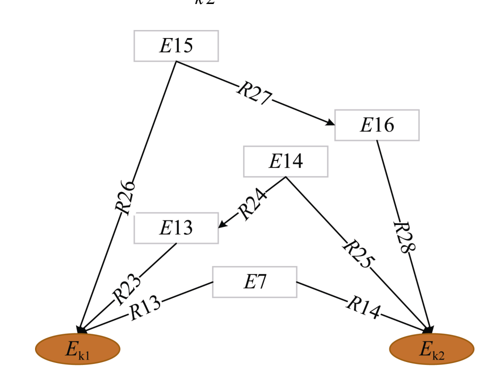


图 3　最近公共祖先节点关联示意图

(4) 最近公共子孙节点关联; 在关联数据中节点间不仅存在最近公共祖先节点, 也存在最近公共子孙节点。关联数据中, 两个主题词节点通过最短的属性关系链向同一个节点, 那么这个节点就被称作最近公共子孙节点(Lowest Common Descendant, LCD), 存在最近公共子孙节点的关联路径被定义为最近公共子孙节点关联(Lowest Common Descendant Relation, LCDR)。如图 4 所示, 当关联路径长度为 2 时存在

$E_{k1} \xrightarrow{R15} E8(LCD) \xleftarrow{R16} E_{k2}$ 一种 LCDR。当关联路径长度为 3 时，存在 $E_{k1} \xrightarrow{R17} E9 \xrightarrow{R18} E10(LCD) \xleftarrow{R19} E_{k2}$ 和 $E_{k1} \xrightarrow{R20} E12(LCD) \xleftarrow{R21} E11 \xleftarrow{R22} E_{k2}$ 等两种 LCDR。

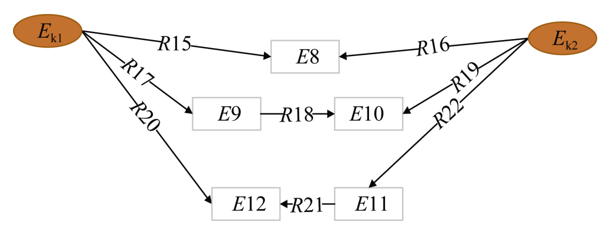


图 4　最近公共子孙节点关联示意图

3.3　关联路径的重要性评价

主题词节点间关联路径数量庞大，无法一一分析，且并非所有关联路径都具有揭示的价值。因此，对关联路径的重要性进行评价是实现类簇实体间语义关系揭示的重要工作，具体包括：实体属性重要性评价、实体节点重要性评价以及实体间路径综合重要性评价。

(1) 实体属性重要性评价

目前，常见的基于关联数据的属性重要性评价指标方法主要包括：基于信息熵的属性重要性评价、基于属性频率的属性重要性评价、基于关联节点的属性重要性评价以及基于 TF-IDF 的属性重要性评价。

①基于信息论

Meymandpour 等[23]提出基于信息论的关联数据信息量衡量方法。信息论利用不确定性度量信息的大小，因此单个关联属性 P 的信息量可以表示为其出现概率的负对数，计算公式为：$I(R) = -\log Pr(P)$，其中 $Pr(P)$表示属性 P 在整个数据集中的出现概率，计算方式为属性 P 出现频次除以关联数据集中属性总频次，公式中对数一般取 2 为底，单位为比特。

②基于属性频率

Kasneci 等[24]基于属性频率构建了信息量计算方法 MING，MING 给出了节点 i 到节点 j 的关联关系 r 的权重计算方法，计算公式为：$Wij=\dfrac{N(i,r,j)}{N(*,r,j)}$，其中 $N(i, r, j)$为实例(i, r, j)的数量，$N(*, r, j)$为所有经由关联关系 r 到达节点 j 的实例数量。Balmin 等[25]将基于属性频率的方法与人工分配权重的方式相结合，在计算属性权重时先根据经验预先给定每种属性分配一定的权重，然后根据关联关系实例数量等比例均分给定权重。Nie 等[26]提出对象排序算法 PopRank 中对关联属性的权重计算也是基于同样的思路。

③基于关联节点

Ng 等[27]提出基于属性所关联的节点计算属性权重的 MultiRank 算法。在 MultiRank 算法中，属性关系的重要性由该属性关系所关联的两个节点(即关联数据中的主语和对象)的重要性得分乘积计算。

④基于 TF-IDF

关联数据集中属性关系的分布往往是偏斜的，不同的属性关系的频率数量级差异十分悬殊。为解决基于属性频次的重要性评价方法的不足，本文提出基于 TF-IDF 的属性权重计算方法。在关联数据中，一个属性在关联数据图挖掘出的子图(如图 1)中出现的频率越高说明它在区分该子图内属性方面的能力越强(TF)，一个属性在整个关联数据集中出现的频率越高说明它的区分性越低(IDF)。基于 TF-IDF 的关联数据属性权重计算公式可表示为：$W_R = tf_{iR} \times idf_R = tf_{iR} \times \log(N/n_R)$，其中 tf_{iR} 指属性 R 在关联挖掘结果子图出现的次数，idf_j 指属性 R 频次的倒数，N 表示关联数据集中的总关联数，n_j 指 t_j 在整个关联数据集中出现的总次数。

(2) 实体节点重要性评价

基于关联数据的节点评价方法主要包括：信息论法、网络图分析法、张量分解法[28]。

①信息论法

信息论中如果一个事件是由若干个独立的小事件构成，则信息量是这些独立小事件的信息量之和。在关联数据中，节点由若干个关联属性组成，节点自身信息量为其关联属性的信息量之和。

②网络图分析法

关联数据网络中的节点和属性类似于 Web 中的网页和超链接，因此传统的网络图分析算法 PageRank、HITS 经过一定的调整也可以应用到 LOD 中。在关联数据中，以某个节点为核心时，可以通过综合考虑核心节点的每个相邻节点通过关联关系对核心节点贡献重要性，形成核心节点的总体的重要性，以此评价核心节点的影响力。以待计算权重的节点为核心，其权重的计算公式可表示为：$R(j) = \alpha \sum_{i \in B(j)} R(i) \times Wij + \dfrac{(1-\alpha)}{|E|}$，其中 $B(j)$是所有指向节点 j 的节点集合，其中 Wij 为节点 i 指向节点 j 的关联关系权重；E 为整个关联数据网络中的所有节点；α 为阻尼系数，一般取 0.85。在开始计算时，每个节点的初始重要性值默认是相同的。与之类似，拓展网络图分析算法 HITS 方法也可用于关联数据中节点重要性的评价，Bamba 等[29]基于 HITS 算法通过预定义每个关联关系的权威度权重和中心度权重来计算节点的主观性得分和客观性得分以进行节点重要性排名。

③张量分解法

张量是一种高维数据的组织方法，张量分解指的是张量等高维数据通过 Tucker 和 Parafac 模型等方法将其直接降维成几个更小更简单的子矩阵相乘来表示的过程，其中分解后的小矩阵描述的是分解前原矩阵的重要特性。关联数据

网络中包含大量丰富的语义关系使其可以表示为一个三维张量 T, 同样, 关联数据中的节点、相邻及连接相邻节点的关系也能表示为三维张量。以待计算节点作核心, 可以通过综合考虑核心节点对各主题的权威度形成核心节点的总体权威度, 以此评价核心节点的影响力[30]。

(3) 关联路径综合重要性评价

基于节点间路径越短语义越相关的一般假设, 可利用社会网络分析中的拓展卡茨中心度指标(Katz's Centrality Measure)对一条路径 P 的重要性进行综合计算, 其基本原理[31]是: 假设两个节点间的路径的有效性由已知的常量概率 α 决定, 那么在一个由 k 个节点组成的路径的概率为 α^k。本文在卡茨中心度指标的基础上引入属性的概率, 长度为 N 的关联路径综合的重要性 $Pr(P)$可通过以下公式计算: $Pr(P)=W(R_1)\times W(E_1)\times W(R_2)\times\cdots W(R_N)$。由于属性和节点的重要性评价结果数量级存在差异, 计算关联路径综合重要性前须对属性和节点的重要性评价结果进行归一化处理。常见的归一化算法包括: 线性函数转换、对数函数转换、反正切函数转换和线性与对数函数结合等方法。

4　基于关联数据的类簇语义揭示实现

本文以 Java 语言和 Eclipse 为开发环境, 借助 Jena 和 Virtuoso 等开源工具和 DBpedia(2016-4)关联数据集实现基于关联数据的类簇语义揭示。

4.1　实验数据的选择

调研发现, 相对于其他数据集, DBpedia 数据更为全面和丰富。DBpedia 是基于 Wikipedia、语义 Web 和关联数据技术的创新型知识库, 是文档网向数据网过渡的标志性成果之一。最新的 DBpedia(2016-4)拥有超过 90 亿个 RDF 三元组, 包含 754 个类, 涉及 127 种语言, 仅英文版的 DBpedia 知识库中就描述了超过 600 万个事物(其中 520 万个资源都归类于统一的本体), 包含 150 万人、81 万个地点、13.5 万份音乐作品、10.6 万部电影、27.5 万个组织机构、30.1 万个生物物种及 5 000 多种疾病, 是目前最大的跨领域语义知识库之一。鉴于 DBpedia 丰富的语义关系和资源规模, 本文采用 DBpedia 数据集作为类簇语义发现的基础。同时, 为保证实验的合理和客观, 选择论文《基于共词分析的兽医分子生物学领域研究热点分析及初步展望》[32]中类簇主题词"Cloning"和"PCR"作为语义揭示的对象。

4.2　语义揭示系统框架

为实现基于关联数据的类簇语义揭示, 笔者设计了如图 5 所示的语义揭示系统框架, 该框架分为关联数据图挖掘和语义揭示两部分。

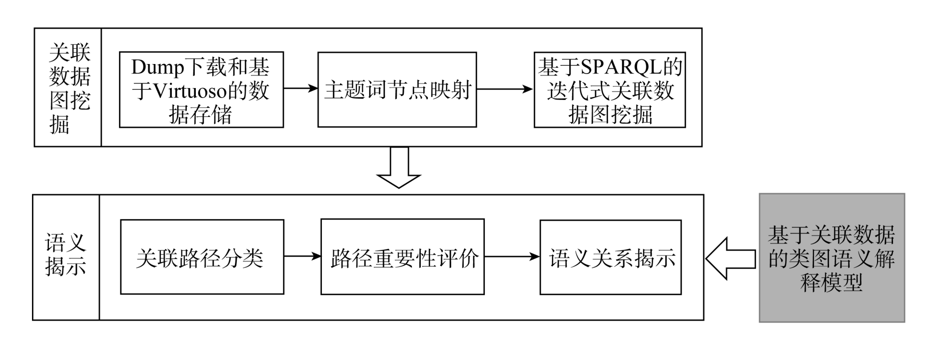


图 5　基于关联数据的类簇语义揭示框架

4.3　关联数据图挖掘

关联数据图挖掘指从关联数据集中发现图 1 所示的主题词节点中关联路径的过程, 分为数据准备和关联数据挖掘两个部分。

(1) 数据准备

以 Dump 下载的方式获取 DBpedia(2016-4)英文版数据集, 并基于 Virtuoso7.2.4 搭建本地 SPARQL 查询。完成关联数据集的获取后, 通过语义浏览器 LodLive 提供的关键词检索服务完成主题词节点的映射, 发现类簇内主题词"Cloning"和"PCR"在 DBpedia 中对应的节点 URI 分别是"http://dbpedia.org/ resource/Cloning"和"http://dbpedia.org/resource/Polymerase_ chain_reaction"。

(2) 关联数据图挖掘

本文在借鉴相关挖掘技术基础上提出基于迭代式 SPARQL 查询的关联数据图挖掘方法, 基本原理是通过 SPARQL 检索迭代查找的方法实现节点间最短关联

路径的发现，查找策略是从长度为 1 的路径开始逐渐增加。以主题词节点"http://dbpedia.org/resource/Cloning"和"http://dbpedia.org/resource/Polymerase_chain_reaction"为初始节点，设定最大挖掘路径长度为 3 在 DBpedia(2016-4)关联数据集中进行挖掘，共计发现 9 480 条关联路径，其中长度为 1 的关联路径 1 条、长度为 2 的关联路径为 72 条、长度为 3 的关联路径 9 407 条，如图 6 所示。

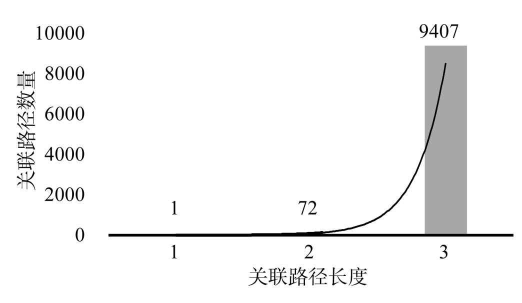


图 6 关联路径数量随路径长度变化趋势

4.4 语义揭示

语义揭示模块指将基于关联数据的类簇语义揭示模型应用于关联数据图挖掘的结果，分为关联路径分类、重要性指标计算和语义关系揭示等三个部分。

(1) 关联路径分类

挖掘关联路径后，根据模型对 4 种关联路径类型的定义，对发现的 9 480 条关联路径进行分类。其中，属于直接关联的有 1 条(0.01%)，属于间接关联的有 1 076 条(11.35%)，属于最近公共祖先节点关联的有 3 847 条(40.58%)，属于公共祖先节点关联的关联路径有 4 556 条(48.05%)。

(2) 路径重要性指标计算

基于现有数据的情况和可行性，笔者基于信息论计算属性和节点重要性指标，并据此评价关联路径的综合重要性。计算过程如下：首先，基于 SPARQL 动态获取属性频次、节点频次等评价所需数据，然后利用第 3 节的方法对属性重要性和节点重要性分别进行计算，最后根据调整 Min-Max 归一化处理后的属性和节点重要性指标计算结果评价关联路径的综合重要性，调整 Min-Max 归一化方法函数表示为 $x^* = 0.001 + \frac{x - Max}{Max - Min} \times 0.999$。路径的综合重要性指标的计算结果如表 1 所示，其中"<*>"表示节点，"$\xrightarrow{*}$"表示属性。

表 1 部分关联路径综合重要性指标计算结果

关联路径	重要性指标	类型
$\{<Cloning>\xrightarrow{\text{http://dbpedia.org/ontology/wikiPageWikiLink}}<PCR>\}$	0.001	DR
$\{<Cloning>\xrightarrow{\text{wikiPageWikiLink}}<Cloning_vector>\xrightarrow{\text{wikiPageWikiLink}}<PCR>\}$	0.00000072	IR
$\{<Cloning>\xrightarrow{\text{wikiPageWikiLink}}<Bisulfite_sequencing>\xrightarrow{\text{wikiPageWikiLink}}<PCR>\}$	0.00000072	IR
$\{<Cloning>\xrightarrow{\text{http://www.w3.org/2004/02/skos/core\#broader}}<Category{:}Cloning>\xrightarrow{\text{http://www.w3.org/2004/02/skos/core\#broader}}<Category{:}Biotechnology>\xleftarrow{\text{http://purlorg/dc/terms/subject}}<PCR>\}$	0.00118999	LCAR
$\{<Cloning>\xrightarrow{\text{wikiPageWikiLink}}<Molecular_cloning>\xrightarrow{\text{http://purlorg/dc/terms/subject}}<Category{:}Molecular_biology>\xleftarrow{\text{http://purlorg/dc/terms/subject}}<PCR>\}$	0.000260651	LCAR
$\{<Cloning>\xleftarrow{\text{http://purlorg/dc/terms/subject}}<Category{:}Molecular_biology>\xrightarrow{\text{http://purlorg/dc/terms/subject}}<PCR>\}$	0.00720822	LCDR
$\{<Cloning>\xleftarrow{\text{rdf:type}}<http://dbpedia.org/dbtax/Technique>\xrightarrow{\text{rdf:type}}<PCR>\}$	0.00139680	LCDR

(3) 语义关系揭示

在揭示关联路径所表达的语义关系前，需分析关联数据中属性关系的语义含义。如表 2 所示，通过 SPARQL 检索获取并分析 DBpedia 中的高频属性，将关联数据的语义关系界定为：等同关系(包含同义和近义)、上下位类关系(属种关系)、整部关系、类与实例的关系以及相关关系(除上述4种关系的其他所有关系)等 5 种基本语义关系，并基于这 5 种基本语义关系分析关联路径所蕴含的语义关系。

①直接关联的语义关系揭示

主题词节点"Cloning"和"PCR"间存在 1 条直接关联：$\{<Cloning>\xrightarrow{\text{wikiPageWikiLink}}<PCR>\}$ ($\xrightarrow{\text{wikiPageWikiLink}}$ 代表属性"http://dbpedia.org/ontology/wikiPageWikiLink")，它所表示的语义关系为：主题词"Cloning"和"PCR"具有相关关系。

表 2　DBpedia 高频属性(部分)

序号	属性	出现频次	含义	语义关系
1	http://dbpedia.org/ontology/wikiPageWikiLink	172 300 574	对应 Wikipedia 的链接信息	相关关系
2	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	66 418 990	资源的标签信息	类和实例关系
3	http://www.w3.org/2002/07/owl#sameAs	40 637 907	指向同义资源	等同关系
4	http://dbpedia.org/property/wikiPageUsesTemplate	36 772 939	RDF 抽取所用模版信息	相关关系
5	http://dbpedia.org/ontology/wikiPageWikiLinkText	23 809 294	Wikipedia 超链接的文本信息	相关关系
6	http://purl.org/dc/terms/subject	22 673 220	资源的主题信息	类和实例关系

②间接关联的语义关系揭示

主题词节点“Cloning”和“PCR”间存在 1 076 条间接关联，其中综合重要性最高的关联路径为：{<Cloning> $\xrightarrow{\text{wikiPageWikiLink}}$ <Cloning_vector> $\xrightarrow{\text{wikiPageWikiLink}}$ <PCR>}，它表示节点“Cloning_vector”与主题词“Cloning”和“PCR”同时具有相关关系。除此之外，实验还发现“DNA”“DNA_sequencing”“DNA_profiling”和“Molecular_cloning”等多个资源也与主题词“Cloning”和“PCR”同时具有相关关系。

③LCAR 的语义关系揭示

主题词“Cloning”和“PCR”对应节点间存在最近公共祖先节点关联 3 847 条，其中综合重要性最高为：{<Cloning> $\xrightarrow{\text{http://www.w3.org/2004/02/skos/core\#broader}}$ <Category: Cloning> $\xrightarrow{\text{http://www.w3.org/2004/02/skos/core\#broader}}$ <Category: Biotechnology> $\xleftarrow{\text{http://purl.org/dc/terms/subject}}$ <PCR>}，它所表示的语义关系为：主题词“Cloning”和“PCR”与类“Biotechnology(生物技术)”都具有上下位类的语义关系，即主题词“Cloning”和“PCR”都是隶属于生物技术类的概念。

④LCDR 的语义关系揭示

主题词节点“Cloning”和“PCR”间存在最近公共祖先节点关联 4 556 条，其中综合重要性最高的为关联路径{<Cloning> $\xleftarrow{\text{http://purl.org/dc/terms/subject}}$ <Category:Molecular_biology> $\xrightarrow{\text{http://purl.org/dc/terms/subject}}$ <PCR>}，它所表达的语义关系为：主题词“Cloning”和“PCR”都与类“Molecular_biology(分子生物学)”具有类和实例关系，即主题词“Cloning”和“PCR”都是隶属于分子生物学类的概念。除此之外，关联路径{<Cloning> $\xleftarrow{\text{rdf:type}}$ <http://dbpedia.org/dbtax/Technique> $\xrightarrow{\text{rdf:type}}$ <PCR>}表示实主题词“Cloning”和“PCR”同时与类“<http://dbpedia.org/dbtax/Technique>”具有类和实例的语义关系，即“Cloning”和“PCR”都是同种技术。

4.5　实验结果分析

笔者对重要性指标排名前 300 的关联路径进行分析，结果显示由于关联数据不完整等质量问题导致的无价值关联路径有 136 条，其余 164 条有语义价值的关联路径中，LCDR 有 106 条(64.6%)，LCAR 有 54 条(32.9%)，IR 有 3 条(1.8%)，DR 有 1 条(0.6%)，可以发现 LCAR 和 LCDR 对类簇的语义揭示最为重要。对 164 条关联路径所揭示的语义关系类型进行分析发现，相关关系以 92.7%(152 条)占据绝对优势，其次是类与实例关系占比 4.8%(8 条)，最后是上下位类关系占比 2.4%(4 条)。相关关系占比最高的主要原因是实验所用数据集 DBpedia 抽取自维基百科，存在大量涉及维基百科网页信息的属性，例如属性“http://dbpedia.org/ontology/wikiPageWikiLink”出现 1.7 亿次，占数据集属性总数(6.8 亿)约四分之一，这些对应相关关系的属性大量存在，造成语义揭示结果中相关关系占比最高。

本实验利用关联数据有效揭示了主题词间的相关关系、类和实例关系以及类和属性关系等多种语义关系，例如：主题词“Cloning”和“PCR”都是隶属于生物技术类的概念、主题词“Cloning”和“PCR”都隶属于分子生物学类的概念、主题词“Cloning”和“PCR”都属于一种技术等。在论文《基于共词分析的兽医分子生物学领域研究热点分析及初步展望》中，专家通过对类簇的人工分析将主题词“Cloning”和“PCR”所属的类簇命名为“克隆技术研究”，与本实验语义结果揭示相一致，证明了基于关联数据的类簇语义关系揭示模型具有可行性和有效性。

实验也存在一些不足，首先仅基于单一的 DBpedia 英文版关联数据集对模型进行实验验证，揭示出的语义关系类型局限为相关关系、类与实例关系以及上下位类关系等三种。另外，关联数据资源存在数据不完整、数据重复和数据不一致等质量问题也对语义揭示的精确度造成一定影响。

5　结　语

本文提出利用关联数据揭示类簇内主题词间的语义关系并通过实证验证了模型和方法的有效性，弥补

了传统类簇分析在语义关系揭示方面的不足，为类簇语义关系的揭示提供了一种新的思路。相较于其他语料库，关联数据具有语义资源覆盖广和结构化程度高的双重优势，快速发展的LOD资源保证了绝大多数领域的类簇可以得到有效语义揭示。本研究主要存在以下两个方面的不足：局限于单一数据集的语义揭示以及关联数据质量对语义揭示结果造成的影响。后续研究中，将对基于更多关联数据资源的类簇语义揭示进行研究，同时改进关联路径的重要性评价指标，克服关联数据质量对语义揭示结果的影响。

参考文献：

[1] 钟伟金，李佳．共词分析法研究(一)——共词分析的过程与方式 [J]. 情报杂志, 2008, 27(5): 70-72. (Zhong Weijin, Li Jia. The Research of Co-word Analysis (1) ———The Process and Methods of Co-word Analysis [J]. Journal of Intelligence, 2008, 27(5): 70-72.)

[2] 张树良，冷伏海．基于文献的知识发现的应用进展研究 [J]. 情报学报, 2006, 25(6): 700-712. (Zhang Shuliang, Leng Fuhai. Study on the Applicational Development of Literature-based Knowledge Discovery [J]. Journal of the China Society for Scientific and Technical Information, 2006, 25(6): 700-712.)

[3] 张晗，任志国，张健，等．基于主题词关联规则的医学文本数据库数据挖掘的尝试 [J]. 医学信息学杂志，2008, 29(1): 32-35. (Zhang Han, Ren Zhiguo, Zhang Jian, et al. Study on the Data Mining in Medical Text Database Based on Keywords Association Rules [J]. Journal of Medical Informatics, 2008, 29(1): 32-35.)

[4] 张晗，崔雷．生物信息学的共词分析研究 [J]. 情报学报，2003, 22(5): 613-617. (Zhang Han, Cui Lei. Study of Bioinformatics through Co-word Analysis[J]. Journal of the China Society for Scientific and Technical Information, 2003, 22(5): 613-617.)

[5] Cimino J J, Barnett G O. Automatic Knowledge Acquisition from Medline [J]. Methods of Information in Medicine, 1993, 32(2): 120-130.

[6] 刘明岩．面向语义关系发现的文本挖掘研究[D]. 南京：南京理工大学, 2010. (Liu Mingyan. Research of Text Mining About Semantic Relation Recognition[D]. Nanjing: Nanjing University of Science and Technology, 2010.)

[7] 张小刚．基于中医药本体的语义关系发现及验证方法[D]. 杭州：浙江大学，2010. (Zhang Xiaogang. Traditional Chinese Medical Ontology Based Semantic Relation Discovering and Verification [D]. Hangzhou: Zhejiang University, 2010.)

[8] 魏来．基于在线词表的Folksonomy语义关联识别方法研究 [J]. 图书情报工作，2011, 55(5): 104-108. (Wei Lai. Research of Folksonomy Semantic Association Method Based on Online Thesaurus [J]. Library and Information Service, 2011, 55(5): 104-108.)

[9] Tiddi I, D'Aquin M, Motta E. Dedalo: Looking for Clusters Explanations in a Labyrinth of Linked Data [M]. Springer International Publishing, 2014.

[10] Taheriyan M, Knoblock C A, Szekely P, et al. Leveraging Linked Data to Infer Semantic Relations Within Structured Sources[C]// Proceedings of the 6th International Workshop on Consuming Linked Data (COLD). 2015.

[11] 李楠，张学福．基于关联数据的知识发现模型研究 [J]. 图书馆学研究, 2013, 1: 73-77. (Li Nan, Zhang Xuefu. Research on Knowledge Discovery Based on Linked Data [J]. Researches in Library Science, 2013, 1: 73-77.)

[12] 李俊，黄春毅．关联数据的知识发现研究 [J]. 情报科学，2013, 31(3): 79-84. (Li Jun, Huang Chunyi. Knowledge Discovery in Linked Data [J]. Information Science, 2013, 31(3): 79-84.)

[13] 高劲松，李迎迎，刘龙，等．基于关联数据的知识发现模型构建研究[J]. 情报科学, 2016, 34(6): 10-13. (Gao Jinsong, Li Yingying, Liu Long, et al. Research on Construction of the Knowledge Discovery Model Based on Linked Data [J]. Information Science, 2016, 34(6): 10-13.)

[14] 宋丽娜．关联数据环境下基于知识地图的隐性知识发现模型研究[D]. 武汉：华中师范大学，2014. (Song Lina. Research on Model of Knowledge Discovery Based on Knowledge Map Under the Environment of Linked Data [D]. Wuhan: Central China Normal University, 2014.)

[15] 刘龙．基于关联数据的知识发现过程模型研究 [D]. 武汉：华中师范大学，2014. (Liu Long. Research on Model of Knowledge Discovery Process Based on Linked Data [D]. Wuhan: Central China Normal University, 2014.)

[16] Narasimha V, Kappara P, Ichise R, et al. LiDDM: A Data Mining System for Linked Data [C]// Proceedings of the 2011 Linked Data on the Web. 2011.

[17] Paulheim H, Fürnkranz J. Unsupervised Generation of Data Mining Features from Linked Open Data[C]//Proceedings of the International Conference on Web Intelligence, Mining and Semantics. 2012.

[18] Ramezani R, Saraee M, Nematbakhsh M A. Finding

Association Rules in Linked Data, A Centralization Approach[C]//Proceedings of the 21st Iranian Conference on Electrical Engineering. 2013.

[19] Personeni G, Daget S, Bonnet C, et al. Mining Linked Open Data: A Case Study with Genes Responsible for Intellectual Disability [M]. Springer International Publishing, 2014.

[20] Jiang X, Zhang X, Gao F, et al. Graph Compression Strategies for Instance-Focused Semantic Mining [C]//Proceedings of the 7th Chinese Semantic Web Symposium on Linked Data and Knowledge Graph. 2013.

[21] Li K, Gao J, Guo S, et al. LRBM: A Restricted Boltzmann Machine Based Approach for Representation Learning on Linked Data[C]// Proceedings of the IEEE International Conference on Data Mining. 2014.

[22] 夏立新, 谭荧. LOD 的网络结构分析与可视化 [J]. 现代图书情报技术, 2016(1): 65-72. (Xia Lixin, Tan Ying. Analysis and Visualization of the LOD Network Structure [J]. New Technology of Library and Information Service, 2016(1): 65-72.)

[23] Meymandpour R, Davis J G. Linked Data Informativeness [M]. Springer Berlin Heidelberg, 2013.

[24] Kasneci G, Elbassuoni S, Weikum G. MING: Mining Informative Entity-Relationship Subgraphs [C]// Proceedings of the 18th ACM Conference on Information and Knowledge Management. 2009.

[25] Balmin A, Hristidis V, Papakonstantinou Y. Objectrank: Authority-based Keyword Search in Databases[C]// Proceedings of the 30th International Conference on Very Large Data Bases.2004.

[26] Nie Z, Zhang Y, Wen J R, et al. Object-level Ranking: Bringing Order to Web Objects[C]//Proceedings of the 2005 International Conference on World Wide Web. 2005.

[27] Ng M K P, Li X T, Ye Y M. MultiRank: Co-ranking for Objects and Relations in Multi-relational Data [C]// Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011.

[28] 蒋世银, 李春旺. 基于关联数据的科研机构评价研究述评 [J]. 情报理论与实践, 2015, 38(2): 136-140. (Jiang Shiyin, Li Chunwang. Review on the Evaluation of Scientific Research Institution Based on Linked Data [J]. Information Studies: Theory & Application, 2015, 38(2): 136-140.)

[29] Bamba B, Mukherjea S. Utilizing Resource Importance for Ranking Semantic Web Query Results[C]//Proceedings of the 2nd International Conference on Semantic Web and Databases. 2004.

[30] Franz T, Schultz A, Sizov S, et al. TripleRank: Ranking Semantic Web Data by Tensor Decomposition[C]// Proceedings of the International Semantic Web Conference. 2009.

[31] Hulpus I, Prangnawarat N, Hayes C. Path-Based Semantic Relatedness on Linked Data and Its Use to Word and Entity Disambiguation[C]// Proceedings of the International Semantic Web Conference. 2015.

[32] 岳阳, 孙静, 石达友, 等. 基于共词分析的兽医分子生物学领域研究热点分析及初步展望 [J]. 广东畜牧兽医科技, 2015, 40(2): 1-4. (Yue Yang, Sun Jing, Shi Dayou, et al. Interpretation and Preliminary Outlook of the Research Focus in Veterinary Molecular Biology Based on the Co-word Analysis [J]. Guangdong Journal of Animal and Veterinary Science, 2015, 40(2): 1-4.)

作者贡献声明:

崔家旺: 文献搜集, 程序设计, 论文撰写;
李春旺: 提出研究思路, 审阅、修改论文。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: cuijiawang@mail.las.ac.cn。

[1] 崔家旺. 关联数据挖掘_9480.xls. 实验数据集.

收稿日期: 2017-02-16
收修改稿日期: 2017-04-11

Identifying Semantic Relations of Clusters Based on Linked Data

Cui Jiawang[1, 2] Li Chunwang[1]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [**Objective**] This paper introduces a model to identify the semantic relations for the co-word analysis results based on linked data. [**Methods**] First, we used Google Scholar, Springer and CNKI to retrieve the literature of the related research. Then, we analyzed the clusters relations of them. Finally, we constructed and examined the semantic relation model for clusters based on the linked data graph structure. [**Results**] The linked data helped us effectively explore the potential semantic relations among keywords. [**Limitations**] Due to the limits of the collected linked data, we only identified some sematic relationship, such as hierarchical, simple relavent, as well as classes-instance ones. More research is needed to improve the quality of linked data. [**Conclusions**] The proposed model could successfully discover the semantic relations among keywords, which help us get more insights from the cluster analysis.

Keywords: Linked Data Co-word Cluster Analysis Cluster Semantic Relations Revealing Model

NISO 发布《标准标签套件》草稿版以征求公众意见

美国国家信息标准化组织(NISO)于近日宣布发布 NISO Z39.102-201x 草案版本, 即《STS: 标准标签套件》(STS: Standards Tag Suite), 以征求公众意见。STS 提供了一种通用的 XML 格式, 标准开发人员、发布商和分销商都可以使用它来发布和交换标准的全文内容和元数据。在草案版本的意见得到解决, 并得到了 NISO 表决委员会和美国国家标准研究所的同意之后, 这一标准将会正式发布, 预计会在今年秋季。

NISO STS 工作组联席主席 Robert Wheeler 说: “在 STS 之前, 有几个 DTD 用于标记标准类型的信息, 这种变化阻碍了跨标准的互操作性, 并且阻碍了组织之间的协作。所以, 各协会、标准制定组织和政府实体一起, 共同创建了这一新工作。该工作是建立在出版商目前正在使用的 ANSI / NISO Z39.96-2015, 即《JATS: 期刊文章标签套件》和标准化国际组织(ISO)的 STS 版本之上。”

JATS 的用户将能够立即熟悉起 STS 模型。NISO STS 工作组联席主席 Bruce Rosenblum 在最近的 STS 电话会议讨论中解释: “在许多方面, 文章的内容与标准内容非常相似, 核心结构部分是相同的, 尽管元数据不同。这个草案是过去 18 个月参与这项工作的两个小组成员以及指导和技术工作组所做的巨大努力的重要里程碑。” NISO 执行董事 Todd Carpenter 也赞赏 JATS 与 STS 之间的协同作用, 他说: “许多标准出版协会都有强大的期刊系统。让这些系统保持一致对这些协会来说就是胜利, 我们期望这两个标准今后也能不断改进。像我们所有的标准一样, 得到了良好的维护, 以满足不断变化和扩展的需求。”

拟议的标准 PDF 格式可从 http://www.niso.org/ workrooms/sts/获得。

(编译自: http://www.niso.org/news/pr/view?item_key=f74de7db56828abfd977e90c2546babc91fdf27d)

(本刊讯)